10: Independent Proportions (2x2 Crosstabs)

Introduction

This chapter considers the analysis of two-independent proportions. Data are stored in the form of a binary outcome (dependent) variable and a binary group (independent) variable.

Techniques will be demonstrated with data from a food borne outbreak study (data are stored in OSWEGO.SAV) in which 75 people attended a picnic and 46 became ill. We will look at exposure to ice cream as a possible predictor to illness. The dependent variable is `CASE` (gastroenteritis: 1 = yes, 2 = no). The independent variable is `ICECREAM` (vanilla ice cream: 1 = yes, 2 = no). We want to compare the proportion of people in each group that became ill. The *first 5* data records are:

```
CASE  ICECREAM
2     2
1     1
1     1
1     1
2     2
etc.
```

The first step of our analysis is to cross-classify (cross-tabulate) the data to form a **2-by-2 table**, with table cells denoted:

Independent Variable	**Dependent Variable** Yes	No	Total
1	a	b	n_1
2	c	d	n_2
Total	m_1	m_2	N

Our illustrative data shows:

ICECREAM	**CASE** 1	2	Total
1	43	11	54
2	3	18	21
Total	46	29	75

SPSS: Data are cross-tabulated by clicking `Analyze | Descriptive Statistics| Crosstabs.`

Estimation

The (incidence) proportion in Group 1 is:

$$\hat{p}_1 = \frac{a}{n_1}$$

For the illustrative example, $\hat{p}_1$ = 43 / 54 = .7963.

The (incidence) proportion in Group 2 is:

$$\hat{p}_2 = \frac{c}{n_2}$$

For the illustrative data, $\hat{p}_2$ = 3 / 21 = .1429. Notice that the outcome occurred much more frequently in Group 1 than in Group 2.

We may wish to estimate the difference in these proportions with confidence. Let us define the risk difference as:

$$\hat{R}D = \hat{p}_1 - \hat{p}_2$$

For the illustrative example, the risk difference = .7963 - .1429 = .6534.

The standard error of this difference is:

$$se_{\hat{p}_1-\hat{p}_2} = \sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}$$

where $\hat{p} = \dfrac{a+c}{n_1+n_2}$. For the illustrative data, $\hat{p} = \dfrac{43+3}{54+21}$ =.6133 and

$se_{\hat{p}_1-\hat{p}_2} = \sqrt{(.6133)(.3867)\left(\frac{1}{54}+\frac{1}{21}\right)}$ =.1252.

When the sample is large (at least 5 cases per group), we use the following formula to calculate a 95% confidence interval for the risk difference:

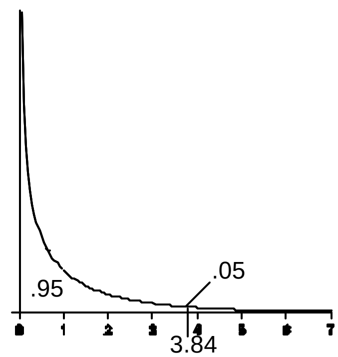
$$\hat{R}D \pm (1.96)(se_{\hat{p}_1-\hat{p}_2})$$

For the illustrative data, a 95% confidence interval for the risk difference = .6534 ± (1.96)(.1252) = .6534 ± .2455 = (.4079, .8989).

Hypothesis Test

The parameters of interest are p_1 (proportion in population 1) and p_2 (proportion in population 2). The two-sided test null and alternative are: H_0: $p_1 = p_2$ vs. H_1: p_1 <u>not</u> = p_2. This is equivalent to H_0: "no association" vs. H_1: "association."

Let us use a **chi-square (c^2) statistic** to perform this test. Chi-square distributions are asymmetrical with long right tails. A chi-square distribution with 1 degree of freedom is shown in the figure to the right. Notice that the 95th percentile on this distribution is equal to 3.84. Let us use the notation $\chi^2_{df,p}$ to denote the pth percentile on a chi-square distribution with *df* degrees of freedom. For example, $\chi^2_{1,.95} = 3.84$. Other chi-square percentiles are found in Appendix 4.



This test statistic is based on a comparison of **observed frequencies (O_i)** to **expected frequencies (E_i)**. The observed frequencies are counts in the sample (see page 1). Expected frequencies are *hypothetical counts,* assuming the null hypothesis were true. These are calculated:

$$E_i = \frac{\text{row total} \times \text{column total}}{\text{total sample size}}$$

For the illustrative data, the expected frequencies are:

	CASE		
ICE CREAM	1	2	Total
1	54 × 46 / 75 = 33.12	54 × 29 / 75 = 20.88	54
2	21 × 46 / 75 = 12.88	21 × 29 / 75 = 8.12	21
Total	46	29	75

Chi-square statistics should *not* be used when an expected frequency is less than 5. Notice that expected frequencies in the above table all exceed 5.

Pearson's ("uncorrected") chi-square test statistic is:

$$ {}^{2}_{\text{stat}} = \sum \frac{(O_i - E_i)^2}{E_i} $$

For the illustrative data, χ^2_{stat} = [(43 ! 33.12)² / 33.12] + [(11 ! 20.88)² / 20.88] + [(3 ! 12.88)² / 12.88] + [(18 ! 8.12)² / 8.12] = 2.95 + 4.68 + 7.58 + 12.02 = 27.23. Under the null hypothesis, the test statistic has (r -1)(c -1) degree of freedom, where r represents the number of rows in the table and c represents the number of columns. For 2x2 tables, df = (2 - 1)(2 - 1) = 1. The p value is the area under the curve in the right tail of the chi-square statistic on the χ^2_{df} distribution. For the illustrative example, $p < .001$. Therefore, the association is significant.

SPSS: Click Analyze | Descriptive Statistics | Crosstabs | Options button: Chi-square.